***Use this word document to describe all your steps after each problem. Show all commands & screen pages and describe your steps. Points are deducted for missing steps.***

Describe all steps for each problem set and capture screen shots for each step with a JPG image within this document. Do not forget to put your name on the top of this file for your submitted homework. Also, please replace \_ALL\_ text in red with your information! Please include in your MS Word document only relevant portions of the console output or output files. Sometime either console output or the result file is too long and including it into the MS Word document makes that document too hard to read. PLEASE DO NOT EMBED files into your MS Word document. You are not obliged to use Java or Eclipse. You are welcome to use any language and any IDE of your choice.

**Problem 1:**

Remove the header of the attached Samll\_Car\_Data.csv file and then import it into Spark. Randomly select 10% of you data for testing and use remaining data for training. Look initially at horsepower and displacement. Treat displacement as a feature and horsepower as the target variable. Use MLlib linear regression to identify the model for the relationship. Use test data to illustrate accuracy of your ability to predict the relationship. Create a diagram using D3 which presents the model (straight line), original test data and predictions of your analysis. Please label your axes and use different colors for original data and predicted data.

**Total points: 35**

Solution:

**STEP 1: Installation Process**

*Java Installation Confirmation*

Graphical user interface, text, email

Description automatically generated

*Spark Installation Confirmation*

Graphical user interface, application, Word

Description automatically generated

The variables to add are, in my example,

| **Name** | **Value** |
| --- | --- |
| SPARK\_HOME | C:\pratik\spark\spark-2.2.1-bin-hadoop2.7 |
| HADOOP\_HOME | C:\pratik\spark\spark-2.2.1-bin-hadoop2.7 |
| JAVA\_HOME | C:\Program Files\Java\jdk1.8.0\_202 |
| PYSPARK\_DRIVER\_PYTHON | jupyter |
| PYSPARK\_DRIVER\_PYTHON\_OPTS | notebook |

in the same environment variable settings window, look for the Path or PATH variable, click edit and add C:\pratik\spark\spark-2.2.1-bin-hadoop2.7\bin to it.

**Step 2: Find the Spark on the system and start the Spark Session**

We can confirm if the Spark session has been properly started through ‘http:/localhost:4042’. As we can see from the image below that a terminal named ‘test’ has been started on the local system.

Graphical user interface, application, Word

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Here we can say it run successfully. Here when we pre run button after that it show the pop up of window firewall, saying access or cancel, I gave access to but whichever you elect it will not affect pyspark.

Graphical user interface, text, application

Description automatically generated

**Step 3: Load the Data and Analyze the Data through PySpark**

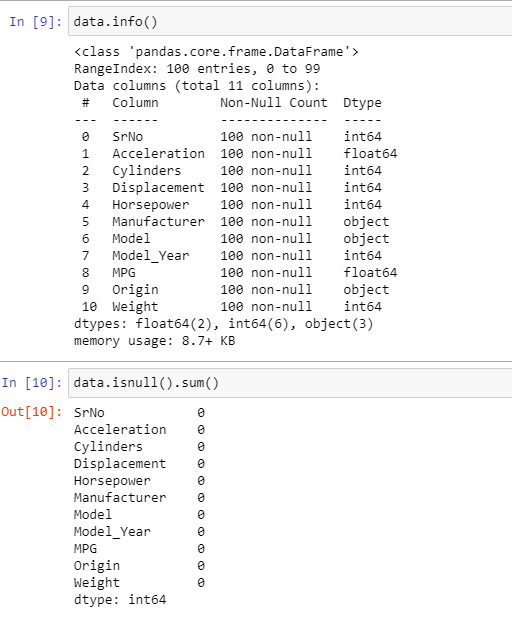
**Graphical user interface, table

Description automatically generated with medium confidence**

**Graphical user interface, text, application

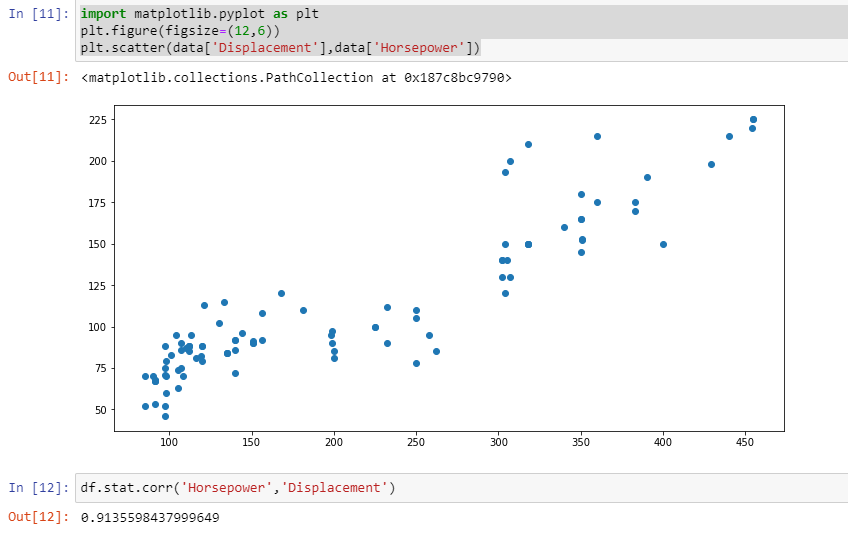
Description automatically generated**

The above image shows the schema of the ‘Small\_Car\_Data’ that we will be using in this assignment. The data contains 11 columns comprising of both categorical and numerical variables.



We can confirm through the above the code that the data is clean and doesnot have missing values. We can now go ahead and perform the tasks that we want to perform

**STEP 4 : Understand the relationship between Horsepower and Displacement**



From the above scatter plot and correlation coefficient, we can clearly see that Horsepower is highly dependent and correlated with Displacement. We can clear establish a linear relationship between the two variables.

**Step 5: Creating the features for Model**

We will use the Vector Assembler function which is available in the Pyspark ML Feature library to convert the feature ‘Displacement’ and target ‘Horsepower’ in Vector form so that we can use MLib to create a model from it

Graphical user interface, table

Description automatically generated

Table

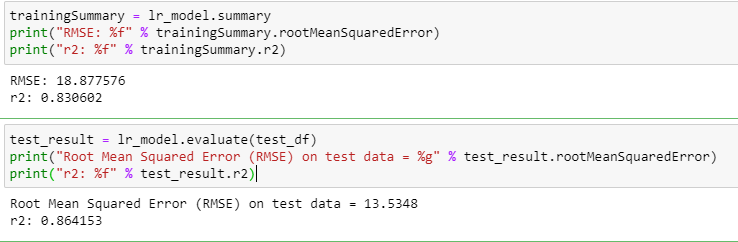
Description automatically generated

**Step 6: Splitting the Data into Training and Testing Data and predicting the re**

Graphical user interface, text, application

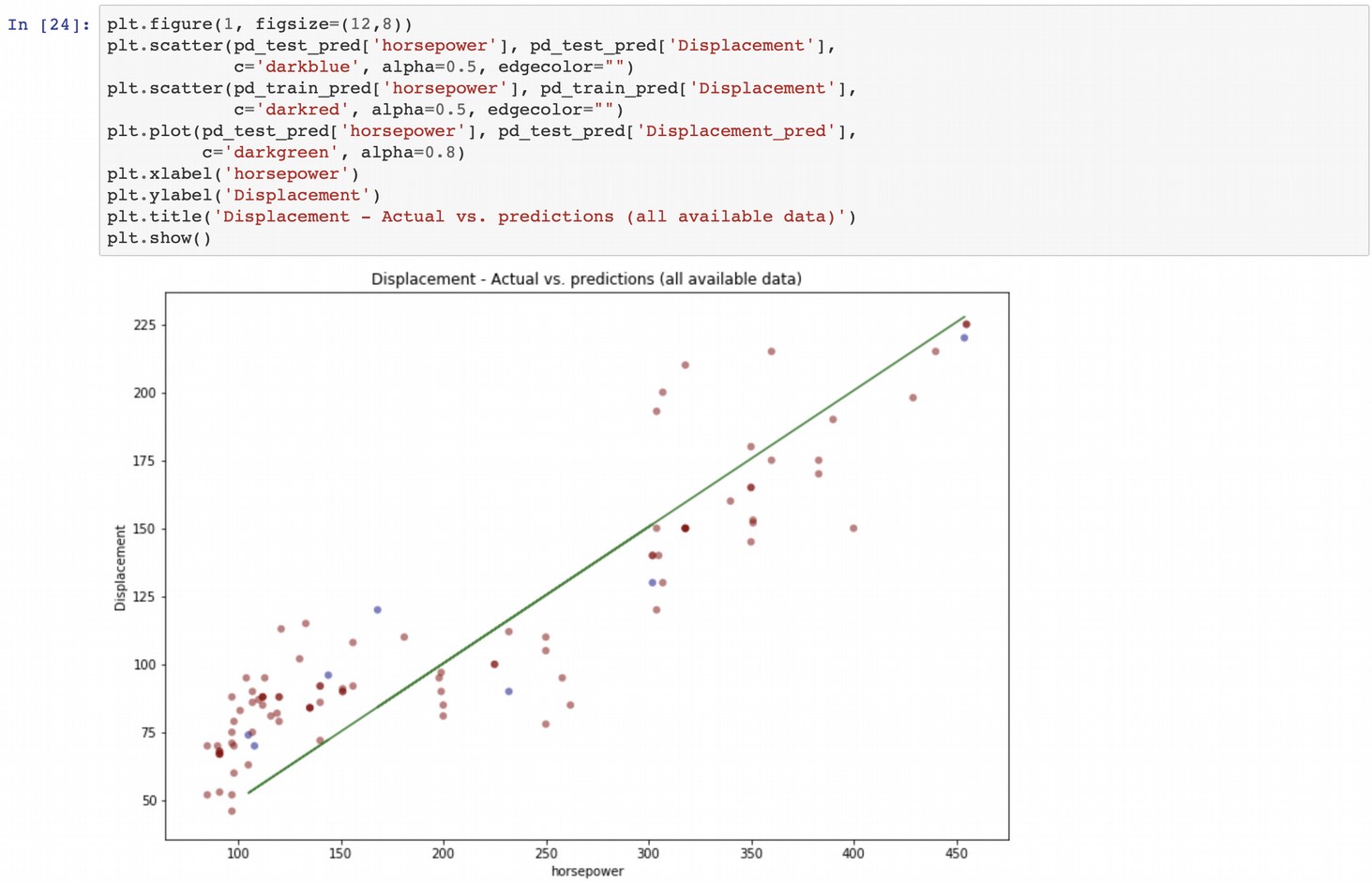
Description automatically generated

**Step 8: Predicting values for Testing Data and Calculating the Accuracy of the Model**



To understand the performance of the linear model created, we will look at the RMSE and R2 value for the Testing Data. As we can see that the R Squared value is 86% which shows that the model that we have created is a very good model with accurate prediction power.

**Step 9: Creating Graph to compare the Actuals vs Predicted values for the Target variable**



As we can see from the scatter plot between the Horsepower and Displacement above that the Actual values are really close to the Predicted Values and the Regression Line is a good predictor for the Horsepower of the car.

**Problem 2:**

Treat: cylinders, displacement, manufacturer, model year, origin and weight as features and use linear regression to predict two target variable: horsepower and acceleration. Please note that some of those are categorical variables. Use test data to assess quality of prediction for both target variables. Which of two target variables is easier to predict, in the sense that predicted values differ less from the original values

**Total points: 35**

**Step 1: Loading and Understanding the data**

We will follow similar steps as we performed in question 1 for creating the linear model by just changing the features. The features for the model that we will create is a combination of categorical and numerical variables.

Graphical user interface, text, application

Description automatically generated

**Step 2: Indexing the Categorical Variables**

As the Linear Regression Model can only handle numerical variables, we would need to encode the categorical variables ‘Model’ and ‘Manufacturer’ that we will be using as features for the prediction of our Target variables.

Graphical user interface, application

Description automatically generated with medium confidence

**Step 3: Converting the Features to Vector**

**Graphical user interface, text, application, email

Description automatically generated**We have created a Vector of the Features that we will be imputing for training the Model – ‘Cylinders’, ‘Displacement’, ‘Model Year’, ‘Weight’, ‘Model’ and Manufacturer’.

**Table

Description automatically generated**

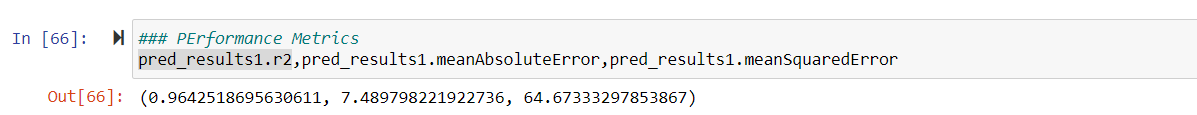
**Step 4: Splitting the Data into Training and Testing data and getting the result:**

**Graphical user interface, text, application, email

Description automatically generated**

We are training the model with ‘Horsepower’ as the Target variable. As we can see from the Coefficients table that Horsepower is highly dependent on the No of Cylinders that are there in the car and is inversely correlated with how old the car is.

**Step 6: Accuracy for the Horsepower model**

****

**R-squared value is 96% here.**

**Step 7: Training the Model & Checking the Accuracy of Acceleration model**

Similarly, we created a model for prediction of Acceleration and then checking the Performance.

**Graphical user interface, text

Description automatically generated**

**Step 9: Evaluating the Best Model**

As the Root Mean Squared Error value of the ‘Accerleration’ prediction model is 1.87which is less than the RMSE test error of the ‘Horsepower’ model, hence, the predicted values of the ‘Acceleration’ prediction model is closer to the Actual values , it is easier and better to predict the Acceleration of the

------------------------------------------------------------------------------------------------------------

Upload your MS Word document named: **EAI6010**\_PratikMantriHW03.docx to the course web site in your folder for Assignment 3. Note: Do NOT include your JPEG files as separate files.

If you have issues with the upload, please notify your instructor! If you are raising issues that might be of interest to all of your colleagues in the class, please use the Discussion Board, Assignment 03 thread on the course site. The Discussion Board is your best friend.

**Reference:**

DataTechNotes. (n.d.). *PySpark Decision Tree Classification Example*. Retrieved October 17, 2021, from <https://www.datatechnotes.com/2021/06/pyspark-decision-tree-classification.html>

*Tutorial 6- Pyspark With Python-Introduction To Pyspark Mlib*. (n.d.). Www.youtube.com. Retrieved October 17, 2021, from <https://www.youtube.com/watch?v=l6dx_0LobsA>

‌

‌